Essential Components Analysis: Next Generation Data Modeling and Dimension Reduction

PI: Kevin Vixie	X-8	.7 FTE	140K/year
Frank Alexander	CCS-3	.3 FTE	80K/year
Marian Anghel	CCS-3	.5 FTE	100K/year
Thomas Asaki	X-8	.3 FTE	60K/year
Bernie Foy	C-PCS	.7 FTE	140K/year
Andy Fraser	X-8	.5 FTE	100K/year
Brad Henderson	NIS-2	.3 FTE	60K/year
Nick Hengartner	D-1	.5 FTE	100K/year
Greg Johnson	CCS-3	.3 FTE	60K/year
Pieter Swart	CNLS	.5 FTE	135K/year
James Theiler	NIS-2	.3 FTE	70K/year
Brendt Wohlberg	T-7	.7 FTE	140K/year
External Collaborators	see [1]		150K/year
Postdocs and students			150K/year
Total Budget	1.48/FY04	1.56M/FY05	1.64M/FY06

Abstract

Though real data lives in very high dimensional spaces, plausible data lies in *typical subsets* (information theoretic term) occupying exponentially small volumes of those spaces. (Picture wispy, slightly fuzzy, low dimensional or thin subsets.) Given an explicit expression for the data subset in the form of a probability distribution, most interesting or important questions can be answered simply. This probability distribution is rarely known, so we resort to approximations informed by advanced geometrical and analytic insights.

The novelty of our approach lies in the combination of (1) domain knowledge, (2) data driven learning, and (3) a unique convergence of insights from a broad range of disciplines, all of which go into the construction of the approximations to these data subsets. Essential components are (roughly) coordinates describing typical sets. They characterize the low dimensional, nonlinear, multiscale nature of these subsets.

The team combines expertise from wavelets (Swart, Wohlberg), dynamical systems (Fraser, Vixie, Swart, Theiler, Anghel), computational methods (Wohlberg, Asaki, Johnson), statistics (Hengartner), remote sensing (Henderson, Foy, Theiler), advanced analysis (Vixie, Swart), dimension reduction (Vixie, Fraser, Hengartner, Wohlberg, Alexander, Anghel, Theiler), and a wide range of physics (everyone). The team already has an extensive history of successful collaboration as well as extensive experience in key components of the proposed work (see [1]).

We will solve four data understanding problems with significant impact for LANL and homeland defense. Specifically, we will improve recognition rates in plume detection, reduce haze left by current state of the art haze removal algorithms, build recognition algorithms for various identification tasks which efficiently factor out invariants, and produce reduced dimensional models of fluid simulations to permit the computationally efficient analysis of uncertainty propagation.

Our tools and results will apply directly to many other LANL data sets arising from real-time sensors, scientific experiments, and large simulations.

Institutional Goals and Objectives Supported by this Proposal: Solutions to many current problems depend pivotally on the efficient analysis of high-dimensional data sets. We propose addressing this need with approaches inspired by insights and experience gained from our various roles in the development of many emerging methods and techniques. The programmatic objectives supported and advanced by the proposed work include (1) an enhanced ability to extract essential intelligence information for homeland defense needs and (2) the development of advanced capabilities for the analysis of high-dimensional datasets from experiments and simulations. In particular, the urgent real-world problems that we will address – remote sensing, detection of weak signals, dynamic model reduction, and identity recognition – are immediately important to homeland defense and laboratory missions. High-dimensional datasets are now commonly produced by modern sensors capable of multiple simultaneous measurements, and they are frequently encountered in image analysis problems involving many pixels of data. The high dimension and complexity of these problems challenge conventional analysis approaches. The work proposed here will result ingeneralized tools that will be easily transferable to other problems of interest to LANL, such as data from fast local probes, bioinformatics and large-scale discrete system simulation.

Science and Technology Objectives: Canonical reduced dimensional descriptions began with Pearson (1901) and Hotelling (1933) and their development of Principal Components Analysis (PCA). The realization in the 1970's and 1980's that random looking behavior could result from low dimensional dynamics drove an effort to understand and measure this low dimensionality. This effort was deeply impacted by the work of Packard, Takens, Fraser, and Theiler (see [2,3]).

Experience suggests that, although the data spaces we encounter are very high dimensional (dimensions $\approx 10^6$ – 10^9 not being unusual), the plausible events only occur in an exponentially small volume of those spaces. This *typical subset* is confined to what can be imagined as wispy, thin subsets not unlike slightly fuzzy, possibly intersecting, low dimensional submanifolds (surfaces). The size of a typical set relative to the data space often quantifies bounds on understanding and task performance. *Essential components* roughly give a coordinate system on this typical subset.

A need for reduced representations of very high dimensional data from many different sources drives the recent revival of interest in dimension reduction. An important aspect of these approaches to dimension reduction continues to be the use of special orthogonal coordinate systems such as the classical PCA bases and the newer wavelet bases. Members of our team have contributed to the development of these wavelet bases for the closely related task of data compression (see *projects/past* at [1]). They have shown that low dimensional models can be competitive and even superior in compression tasks (see *projects/past* at [1]). More recently, another subset of the team used very short term funding to develop a data classification algorithm using nonlinear low dimensional approximations to obtain competitive face recognition results. This work was the subject of an invited talk at a National Academy of Sciences workshop on massive data streams (see *projects/current* at [1]).

At the most basic level, our innovations will allow us to choose an efficient representation of typical subsets. We will apply these methods to real data of great scientific and programmatic interest as outlined below. *More specifically*, we will improve the recognition rates

in plume detection, reduce the residue of haze left by current state of the art haze removal algorithms, build recognition algorithms for various identification tasks which efficiently factor out invariants, and produce reduced dimensional models of fluid simulations to permit computationally efficient uncertainty propagation.

Tasks and Probable Accomplishments: We will focus our analysis on image-based data cubes where the third (or fourth) dimension represents either highly-resolved spectral or temporal information. To ensure wide applicability, we will target four timely applications. The first two address key problems associated with the rapid increase in hyperspectral remote sensing data: weak signal identification in the presence of clutter (plume detection) and material identification through uncertain atmospheric conditions (haze removal). The third addresses the problem of understanding large-scale simulations of complex dynamic phenomena, namely (ECA based) model reduction for efficient analysis of fluid simulation outputs, The fourth, classification in the presence of invariances (face recognition), was chosen as a challenging but comparatively smaller and well-controlled test problem with easy access to large datasets and a large body of published results. We expect the following concrete accomplishments: (1) We will develop, implement and analyze several novel, integrated approaches to adaptive, multi-resolution, nonlinear essential component analysis. We are confident this will lead to revolutionary improvements over state of the art nonlinear principal components analysis. (2) We will use these methods to improve the state of the art in the atmospheric correction of remotely-sensed imagery. (3) We will implement and use an innovative reduced representation framework for fluid simulations and thereby improve prediction, characterization, and sensitivity analysis. (4) We will improve the handling of invariants in computer vision, resulting in improved recognition rates under lighting and pose variation.

The team members combine strong track records in scientific innovation, important programmatic contributions and collaboration both at LANL and with external leaders in science. Additionally, the team already has a history of close and unusually efficient collaboration. For example, our work on a 2002 homeland defense LDRD, resulted in 2 invited talks, a poster and 2 papers in preparation (one of which was invited). For more details see [1].

Impact of the Proposed Research: The immediate impact of the research will be improved performance of plume detection algorithms, reduced haze residuals, better classification mod invariants (e.g. faces) and more closely optimized dynamic model reductions. These results alone have important implications for threat reduction and homeland defense. The work also has great potential to impact scientific understanding of the domains of each of these data sets through the resultant reduction of complexity, enabling us to see the residual and essential complexity much more clearly. Additionally, as a natural outcome of this research we will produce a set of valuable tools for attacking a wide variety of high-dimensional data analysis problems. Moreover, we will build internal and external collaborations which will guarantee the vitality of continued research at the lab. By facilitating an expertise in massive and high-dimensional data analysis, we will guarantee a leading role for LANL in responding to the increasing demands of scientific, programmatic, and homeland security needs.

[1] DDMA website. http://laurel.lanl.gov/~vixie/DDMA.
[2] Fraser and Swinney. *Physical Review A*, 33:1134, 1986.
[3] Theiler et al. *Physica D*, 58:77–94, 1992.